Hindawi
Journal of Sensors
Volume 2021, Article ID 8893661, 13 pages
https://doi.org/10.1155/2021/8893661

Research Article

Empirical Investigation of Multimodal Sensors in Novel Deep Facial Expression Recognition In-the-Wild

Asad Ullah,[1,2] **Jing Wang**,[3] **M. Shahid Anwar**,[3] **Taeg Keun Whangbo**,[2] **and Yaping Zhu**[4]

[1]*Department of Computer Science and Information Technology, Sarhad University of Science and Information Technology, Peshawar 25000, Pakistan*
[2]*Department of Computer Engineering, Gachon University, Seongnam, Republic of Korea*
[3]*School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China*
[4]*School of Information and Communication Engineering, Communication University of China, China*

Correspondence should be addressed to Yaping Zhu; zhu_yaping@hotmail.com

Received 5 May 2020; Revised 22 February 2021; Accepted 5 March 2021; Published 18 March 2021

Academic Editor: Giuseppe Quero

The interest in the facial expression recognition (FER) is increasing day by day due to its practical and potential applications, such as human physiological interaction diagnosis and mental diseases detection. This area has received much attention from the research community in recent years and achieved remarkable results; however, a significant improvement is required in spatial problems. This research work presents a novel framework and proposes an effective and robust solution for FER under an unconstrained environment. Face detection is performed using the supervision of facial attributes. Faceness-Net is used for deep facial part responses for the detection of faces under severe unconstrained variations. In order to improve the generalization problems and avoid insufficient data regime, Deep Convolutional Graphical Adversarial Network (DC-GAN) is utilized. Due to the challenging environmental factors faced in the wild, a large number of noises disrupt feature extraction, thus making it hard to capture ground truth. We leverage different multimodal sensors with a camera that aids in data acquisition, by extracting the features more accurately and improve the overall performance of FER. These intelligent sensors are used to tackle the significant challenges like illumination variance, subject dependence, and head pose. Dual-enhanced capsule network is used which is able to handle the spatial problem. The traditional capsule networks are unable to sufficiently extract the features, as the distance varies greatly between facial features. Therefore, the proposed network is capable of spatial transformation due to action unit aware mechanism and thus forward most desiring features for dynamic routing between capsules. Squashing function is used for the classification function. We have elaborated the effectiveness of our method by validating the results on four popular and versatile databases that outperform all state-of-the-art methods.

1. Introduction

Facial expressions contain the most important nonverbal and rich emotional information in the social communication [1] People communicate with each other through verbal and nonverbal communications [2]. Nonverbal communication is basically conveyed through facial gestures, eye to eye contact, facial expressions, and paralanguage [3]. According to earlier research, in the communication out of 100 percent, the information conveyed by facial expression is 50 percent, 40 percent is through voice, and 8 percent through language. Apart from that, due to the rapid progression in technology, we spend most of the time on electronic devices which carry a variety of software interfaces, which are tense, primitive, and nonverbal. So, facial expression recognition can further improve to have a more natural and intelligent human-machine interaction.

Facial expression recognition is used in various domains like Intelligent Tutoring System (ITS), psychology, human-machine interaction, behavioral science, intelligent transportation, and

interactive games [4]. It can be helpful in monitoring the abnormal expressions in the crowd at public places to avoid any crime. It can also be helpful in the service industry to timely capture the feedback of customers, and it can provide timely treatment of patients by looking at the real-time expressions of the patient at the hospital. According to Ekman and Friesen [5], there are six basic expressions: happiness, surprise, disgust, fear, sadness, and anger (some researchers have termed neutral expression as the seventh expression). These expressions are conveyed almost among all species.

Facial expression recognition is widely studied by various researchers. Despite the available research, robust FER is yet an open and challenging task [6, 7]. However, most of the recognition algorithms did not consider interclass variations caused by the differences in facial attributes of the same individual. Hence, mostly, expression classification is done through facial expression information along with identity-related information [8, 9]. The main drawback it carries is that it affects the overall generalization capability of FER systems, thus resulting in the degradation of the performance on unseen identities [10]. An efficient FER system plays a vital role in the treatment of patients by observing their variable behavior patterns. Happiness expression depicts a healthy and positive mental state, while sad and angry demonstrate an unhealthy mental state. Different mental diseases like autism or anxiety are detected due to the emotional conflicts of a particular patient. Most of the FER systems receive physical signals from camera, but it is also important to observe physiological signals that can be captured with the help of various other sensors. An important application of FER is E-health care; nowadays, almost 0.3 billion people are suffering from depression which can also lead to suicidal tendencies if they are not treated timely and effectively [11]. In general, the mental health treatment observes a lot of barriers like financial cost, social stigma, and shortage of accessible options. Normally, clinical staff holds an interview with a patient for checking symptoms of the depression via verbal and nonverbal indicators. Patients are asked to fill the questionnaire for measurement of depression severity [12]. In order to get timely detection of depression symptoms, an AI-based system will help in entrenching barriers for timely and effective treatment. With the help of multimodal sensors, clinicians can get automatic tools for screening tool of depression by audio, visual, and linguistic signals. Meanwhile, a patient can get his mental status with the help of an AI application in a portable device like the camera.

In this paper, we use a combination of different techniques to develop a robust model. Initially, we implement different preprocessing techniques to fine-tune and remove highly uncorrelated information in the images. Face detection is performed using facial attributes due to the following reasons. (1) The human face is basically having a unique structure, where the most important local facial parts such as eyes, mouth, and nose help us to detect the face in an unconstrained environment. So, the partness map or response map of five different parts is used in the method. (2) The faces observe spatial arrangements like hairs will be above eyes, lips below nose, and so on. Hence, faceness score has been derived from the response configuration. (3) The face hypothesis is performed for the estimation of more accurate face locations. Our contribution is to introduce a special attributes supervision in order to discover facial part responses. We adapt Deep Convolutional Generative Adversarial Network (DC-GAN) for data augmentation. It helps us in the demonstration of realistic data-augmentation and improvement in the generalization performance in the low-data regime. In facial expression recognition, the focus has been shifted towards multimodal sensor data acquisition, adding a reasonable number of intelligent sensors to extract features accurately. This works well in lab-controlled environments and health care facilities. We use multimodal smart sensors for illumination variation pose/head detection and infrared for targeted area/personnel. Considering E-Health care systems, we use sensors for capturing electrical signals such as electrocardiogram (ECG) and electroencephalograph (EEG). A lot of sensors are needed to be added to a system that will be impractical to be used in the real-life scenarios which are most probably be emergency situations such as flooding, fire, earthquakes, and tsunamis. Our proposed method combines normalization techniques to counter environmental noise and unwanted data that hurdles in the accurate emotion detection in the wild.

For accurate and robust FER, feature representation of the facial images is the most important step. A considerable amount of research has been done over local and global feature extraction [13]. Fan et. al [14] suggested a model, i.e., MRE-CNN which aimed to enhance the learning power of the convolutional neural network by considering both the local and global features. Whereas Li et al. [15] introduced the DLP-CNN framework, in which the discrimination power of deep features is enhanced while maximizing the interclass scatter and by preserving the locality closeness. Still, they are unable to find the relative relationship between the local features. Face is composed of a certain structure where every part is having a relative relationship with other part. To address this issue, we propose a method which is capable of spatial transformation due to action unit aware mechanism and thus forward most desiring features for dynamic routing between capsules. Finally, squashing function is used for the Classification purpose. We assess the effectiveness and performance of the introduced model on the Extended Cohn-Kanade, MMI, Oulu-CASIA, and Real-world Affective Faces (RAF) databases. Figure 1 shows some sample images from CK+ database.

The main contributions of this paper are as follows:

(i) For robust face detection, we introduce a special attributes supervision in order to discover facial part responses

(ii) We propose a dual-enhanced capsule network, which is able to extract the effective relationship between the features from different local regions. Spatial information is also encoded due to the knowledge of the probability of an object existence

(iii) The proposed method uses an action unit awareness mechanism which captures more effective and robust information (even subtle muscle motions),

Figure 1: Different facial expression's samples taken from CK+ database.

used for dynamic routing between the capsules, and as a result, provides with much better feature representation

The organization of the next sections is as follows. In Section 2, we provide the problems with the existing methods. In Section 3, we elaborate our novel architecture with the underlying information. Section 4 comprises the results and analysis. Finally, we provide the conclusion of our research and explain the direction for future work in the last section.

2. Related Work

The main goal of FER is to capture the meaningful features which are discriminative and descriptive and invariant to facial variations such as occlusion, illumination, pose, and other identity-related details. There are two main methods available for feature extraction: (1) handcrafted and (2) deep learning-based method. Nowadays, deep learning methods are gaining remarkable results. However, earlier mostly facial expression recognitions were based on handcrafted/human-engineered features such as Histograms of Oriented Gradients (HOG) [16], *n*-dimensional scale-invariant feature transform (n-sift) [17], and Local Phase Quantization (LPQ) [18]. These methods are used for the extraction of the global as well as the detailed information of an individual face. However, the information obtained is from the overall facial region, and it ignores the expression changes in the local regions which contain eyes, nose, and mouth. These methods perform pretty well in the lab-controlled environment where subjects pose expressions under the constant illumination, stable eye gaze, and head pose movement. Existing handcrafted approaches demonstrate comparatively less recognition accuracy. Efforts are exerted for manually extracting the desired discriminating features which are linked to expression changes. Considering in-the-wild scenarios deep learning methods for the robustness of facial expression recognition has been implemented [19–22]. However, deep representation is affected just because every facial attribute of a particular subject carries a hefty number of variations such as gender, ethnicity, and age of the particulars posing expressions. It holds a very big disadvantage, i.e., the generalization capability for any model is highly and negatively affected; as a result of unseen objects, the performance of facial expression recognition is degraded. Although quite a work has been done in the area for the improvement in the performance of FER, still alleviating the influence of intersubject variations is a challenge and open area of research.

Several techniques have been implemented by reducing intraclass variations and by increasing the interclass differences. Hence, it further increases the discriminating property of the features extracted for FER in the real-time scenario [23]. Identity-Aware CNN (IACNN) proposed that by reducing the influence of identity-related information with the use of expression and identity-sensitive contrastive losses, the facial expression recognition performance can be enhanced [24]. The island loss has been proposed for extracting the effective discriminative features for FER [25]. Moreover, in [26], with the use of residue learning, the person-independent expression representation has been learned. However, this technique offered computationally costly; in addition, due to the same intermediate representation used for the generation of neutral images for the same identities, it also was unable to disentangle the expression information from identity information. However, in [24], due to large data expansion caused by the compilation of training data in image pair forms, the effectiveness of contrastive loss is heavily affected [25]. Similarly, in [27], a fixed identity has been proposed for the transfer of facial expressions to fix the influence of identity relative information. The problem still persists with the methods as the efficiency of FER depends on the expression transfer procedure. In short, it has been noticed that FER based on the deep learning methods has outperformed the traditional handcrafted methods. However, there is still a gap in deep learning because very few studies have employed facial depth images in the deep networks as an input. Compared with the existing models, the main goal is to design a network which can be fully adopted for decomposition of the facial region, easy to implement, and is robust.

The shifting of facial expression recognition (FER) from laboratory environment to real-world/wild decreases the efficiency of recognition of correct features from 97 percent to almost 50 percent due to the challenges coming from the environment. These are due to the changing factors such as illumination, head poses, and subject dependence. This problem is mainly solved using multimodal sensors, sensors that are used alongside the main sensor which is the camera. E-

health care system uses electrical devices attached to the subject body or electrodes that need to be injected in the subject's body alongside image/video to make emotion detection more accurate and better facilitate its patients [28]. This makes it unfeasible in the field for an emergency situation. Some of the work is done to achieve accurate results in the wild by improving methods and not the number of sensors in such environment. Challenging light conditions in the wild is taken as a bigger challenge in extracting accurate facial features. Liu et al. [29] devised such a system to handle the illumination problem by using three different classifiers (kernel SVM, logistic regression, and partial least squares), which could be used in the wild. Another work by the same authors [30] shows that this problem can also be handled using the developed method. Recent advancements in technology in sensors and methods, computer vision, speech recognition, deep learning, and related technologies have made emotion detection more accurate and efficient [31, 32]. This still makes it a subject of adding more sensors, which makes sense in its own usage area, but speaking of real-world application in the wild it would be rather feasible to stick to a minimum number of sensors and achieve a method to solve most of the issues if not all.

3. Proposed Method

3.1. Preprocessing. Preprocessing is very important as it aims to capture the meaningful features and align and normalize the most needed visual information conveyed by facial image. Every real-time image is affected by nonlinear facial variations, i.e., varying illumination, difference in the contrast between the foreground and background, and irrelevant head poses. Therefore, to get the maximum possible semantic meanings of the features for further training the deep neural network, we need to perform some preprocessing techniques. This step is used for the elimination of highly uncorrelated data in the image.

3.1.1. Face Detection. Face detection is one of the vital steps in the FER, just due to the excessive background and there is still highly uncorrelated information in the image even taken from few benchmark datasets. As most of the datasets are having almost frontal view and high-resolution images. So Viola and Jones algorithm [33] is used in the most scenarios.

Faceness-net has been used in this paper [34]. A CNN supervised with facial attributes can detect faces even when more than half of the face region is occluded. In addition, the CNN is capable of detecting faces with large pose variation, e.g., profile view without training separate models under different viewpoints. A full image is provided as an input image to the convolutional neural network for the generation of the partness map. The partness map is generated for different facial parts like eyes, nose, and mouth. Facial attributes are further categorized in order to distinguish them from other parts. Just like hairs, it can be blond, black, wavy, straight, etc. So in the next stage, face proposals are much more refined, so that the usefulness of facial attributes are explored for learning optimized and robust face detection. A CNN is trained over uncropped images and is used for obtaining face part detectors without any explicit part supervision. The faceness score is evaluated based on the face part responses and considering the spatial arrangements associated with them. The method is trained on datasets on the following datasets CelebA for face recognition and AFLW for face alignment. After the generation of face proposals, a strong face detector is trained and it outperforms all other methods.

In Figure 2, the face is divided into five important parts, where eyes, nose, and hairs are much more effective as compared to mouth and beard which can be partially occluded. Therefore, the combination of facial parts gives a much better result instead of individual facial parts.

3.1.2. Data Augmentation. As far as deep neural network is chosen for FER, data augmentation is used to produce many better results by providing a large amount of data. It is effective in the generalization capability of the model as many of the publicly available datasets are not large enough to validate the results more efficiently. Large training data yields to a well-trained model.

There are some standard methods of data augmentation like skewing, rotating, shifting changing color scheme, resizing the image, and enhancement of image noise [22]. To automatically learn the augmented data in the low-data setting, we have used Deep Convolutional Generative Adversarial Network (DC-GAN) [35]. It is used for the alleviation of the overfitting problem over the on-the-fly data. The samples provided as input are randomly cropped from all the four sides, and then, the horizontal flip is performed for making a dataset ten times bigger than the original one.

3.2. Multimodal Sensors. In FER systems, the camera is the most commonly used sensor for capturing images and videos of a subject. However, it works well in a lab-controlled environment but its efficiency reduces in the extreme wild real-time scenario. It faces three key challenges: illumination variance, subject dependency, and head pose. The solution to this problem is by adding other dimensions to the feature vector that can be attained from other sensors. To address these problems, we leverage multimodal sensors given below.

3.2.1. Eye Tracker. This is a type of visual sensor used to find some new patterns from a particular section of a face. Among other facial parts, the human eye is considered to be an important part, which helps us to analyse the mental status of an individual such as focus, attention, presence, and consciousness. This can be achieved with the help of an eye-tracking sensor that can provide information about the exact focus of eyes. It assists in the improvement of efficiency in facial expression recognition.

3.2.2. Nonvisual Sensors. There are three main types of nonvisual sensors. The first among the list is voice that is pretty correlated with the body [36]. Most of the FER systems just use visual information instead of taking audio signals with it. Researchers have highlighted the impact of audio signals on the effectiveness of FER. However, proper fusion of audio with visual data is very challenging. In EmotiW challenge 2014, Liu et al. presented expression recognition based on

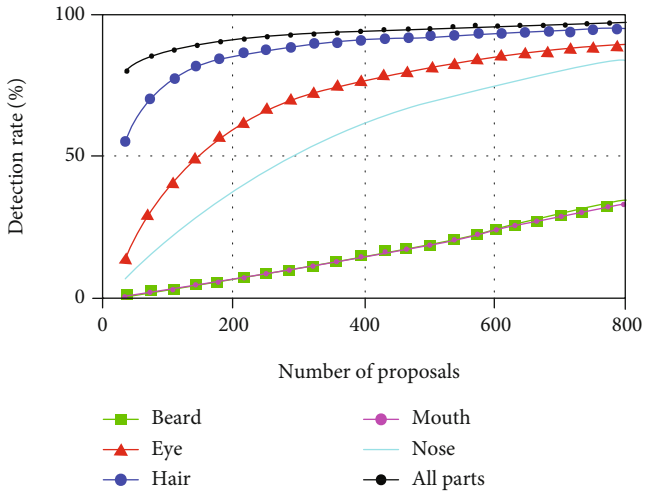


FIGURE 2: Impact of different facial parts to face proposal (individual or as a combination).

the fusion of audio and visual information [29]. We select one-score linear fusion method for taking the optimal features from both visual information and audio signals.

The next among them are sensors for different physiological signals. There are four physiological signals but our primary concern signals are electrocardiogram (ECG) and electroencephalograph (EEG). The former method deals with the heart signals and the later deals with the brain signals. It is basically the acquisition of electrical signals obtained from both of the signals; they are obtained with the attachment of electrodes to the skin. Most of the researchers encouraged EEG because of the robust signals from the brain. However, sometimes, people do not necessarily express their emotions with the facial movement, and illumination variance also affects the recognition process. So, the help of fusion with physiological signals boosts the robustness of the FER system.

The third one is Depth Camera; the in-depth image pixel intensities provide more robust features as compared to RGB images. They have been used in a variety of applications like face detection and body motion recognition. Modified local directional pattern features are obtained from sorting the strengths, among eight directional per-pixel in-depth strengths. These sensors are normally used for the safe privacy of individuals, and it can help us in the noise-affected images. Multichannel features are extracted from RGB facial images, which results in robust FER in an unconstrained environment. Figure 3 shows the fusion of multimodal sensors with pure image/video processing.

3.3. Dual Enhanced Capsule Network (DE-Capsnet). The entire network has been shown in Figure 4, where the model is divided into portions. Firstly, we have to preprocess the images to avoid the uncorrelated information linked to the facial image. Then, we have two modules for further processing. In the first part, the box with the purple dashed line is attention aware of action units and consists of deep convolutional layers for extraction of the enhanced feature maps, and this has been termed as enhancement module 1. In the later part, with the use of dynamic routing, those enhanced feature maps are encoded between capsules, and the process of decoding is done by the fully connected layers (the process has been shown in the green dashed lines). In the end, the squashing function is used for the recognition of facial expressions.

VGG19 is used in the enhancement module 1 just because of the reason that it is very much robust in object classification; meanwhile, it is having a simple architecture too. For a better understanding of the description, each stage is having multiple convolutional layers followed by a max-pooling layer. In the first 2 stages, each stage is having 2 convolutional layers. Whereas in the last 3 stages, each stage is having 3 convolutional layers, respectively. We do not retain the last 3 layers as we have to get the feature maps.

In order to achieve the attention map, we have used the generation method by Li [37]. Furthermore, we have made appropriate adjustments to the datasets used in our work for getting the key facial landmarks. Figure 4 is showing the facial image with having blue facial landmarks, and the attention map has also been showed. Action unit's centres are obtained with key facial points by using scaled distance. In order to make sure that the scales must be the same among all the facial images, the facial images are resized. Hence, for making the shifting distance among the images as much adaptive as much as we can, the measurement reference is used for the shift in distance. To locate the action unit centres of the inner corner distance has been used as scaled distance. For each action unit, the 7 pixels in the nearby area have been taken in the experiments; as a result size of each action unit area is 15×15. Hw is assigned as the higher weight, which are the closest points to the action unit centre.

$$\boldsymbol{H}_{\boldsymbol{w}} = \mathbf{1} - \mathbf{0.07}\boldsymbol{m}_{\boldsymbol{d}}. \tag{1}$$

The manhattan distance is termed as m_d to the action unit centre. Hence, those areas which are having higher values in the attention map correspond to the active areas of action units in facial images; the attention map will further enhance them.

After the generation of attention maps, the maps are further forwarded to stage 3 and stage 4 as shown in Figure 4. The feature maps which are generated after the pooling layer of the second stage are multiplied with the attention map of the first stage and after that being parallel with the convolutional layers of the third stage. Hence, the results obtained after the convolution are added element by element and, then, forwarded to the max-pooling layer of the current stage as an input. A similar operation is done at the fourth stage by jointly combining the convolutional layers with the attention map. Here, we want to explain the reason behind using attention maps; it is just because all areas are not equally important for facial expression recognition.

After the enhancement module 1, we get $512 \times 7 \times 7$ feature maps. For the dynamic routing, the feature maps are further fed between primary capsule layers and face capsule layers. Three fully connected layers are used for decoding and reconstructing the facial image. The nonlinear function, i.e., squashing function is used for facial expression recognition which is defined in Equation (2) as

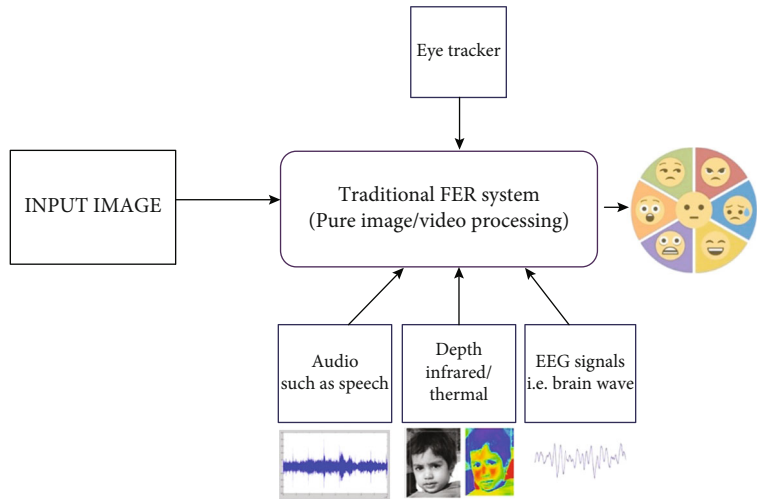


Figure 3: Integration of visual and nonvisual sensors into facial expression recognition.

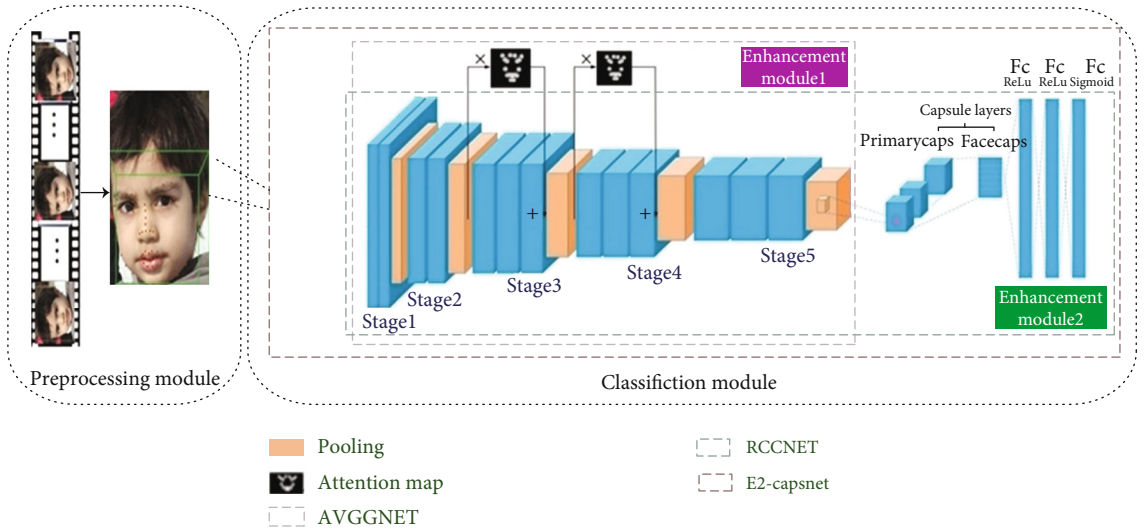


Figure 4: Overview of the proposed method.

$$u_k = \frac{||j_k||^2}{1 + ||j_k||^2} \frac{j_k}{||j_k||}, \tag{2}$$

where k is used for the capsule and u_k and j_k are output and input vectors, respectively. L_m is the minimizing margin loss and L_r is the reconstructing loss used for updating the parameters in the network. Total loss is defined as L_t. Loss function expressions are defined in Equations (3)–(5), respectively.

$$L_m = I_{cc} \max\left(0, b^+ - ||u_{cc}||\right)^2 + \lambda(1 - I_{cc}) \max\left(0, ||u_{cc}|| - b^-\right)^2, \tag{3}$$

$$L_r = (f_c - f)^2, \tag{4}$$

$$L_T = (L_m + 0.0005 L_r)^2, \tag{5}$$

where cc is termed as the classification category, and for that particular category, the indicator function is denoted by I_{cc}. The upper and lower boundaries are represented by b^+ and b^-. The f represents the original image, whereas f_c represents the reconstructed image.

4. Performance Metrics

To evaluate the merits of the proposed method, there are some performance metrics used in the quantitative comparisons of different approaches. There are different methods for comparing recognition rates with other approaches. The difference between the methods depends basically on the difference in division of the training sets and test sets. Facial expression recognition basically comprises of multiclass. 10-fold cross-validation is performed for dividing each expression category in the number of training and test sets. This

method is useful in utilizing all the samples. It is also capable to effectively avoid overfitting and underfitting problems.

A set of evaluation metrics is required to discriminate and obtain an optimal classifier. There are few evaluation metrics implemented in this work to populate the effect of recognition rates. The performance metrics are as follows.

4.1. Accuracy. Accuracy is one of the most widely used metrics in classification problems. The accuracy is achieved as a ratio of correctly predicted samples to all the predictions made. For calculating the average accuracy, we need to average all the accuracies for each category of expression. To calculate the accuracy, the equation is given as follows:

$$\text{Accuracy} = \frac{\text{True}_p + \text{True}_n}{\text{True}_p + \text{True}_n + \text{False}_p + \text{False}_n}, \quad (6)$$

where "p" stands for positive and "n" stands for negative predictions, respectively. Figure 5 represents the average accuracy rates on the aforementioned databases.

4.2. Precision. Precision of each class is the ratio of true positives to true positive and true negative. Precision is highly dependent on the threshold value of classifier. If the threshold value was set high previously, then the precision will increase because the new results may lead to all positive. If the threshold value was low or about right, then the precision will be decreased by lowering of the threshold. It will create more false positives. The equation of the precision is as follows:

$$\textbf{Precision} = \frac{\textbf{True}_p}{\textbf{True}_p + \textbf{False}_p}. \quad (7)$$

4.3. F1-Score. The $F1$-score also called F-measure or balanced F-score is the harmonic mean of the precision and recall for each class. The precision and recall equally contribute in the weighted average of the $F1$-score. The equation is given as follows:

$$\boldsymbol{F1}_{\textbf{score}} = \mathbf{2} * \frac{\textbf{Precision} * \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}, \quad (8)$$

where Recall is the ratio of true positives to all the samples which supposed to be identified as positive.

5. Results and Discussion

We have used the four most popular databases for populating the results. These databases are CK+ [38], MMI [39], Oulu-CASIA [40], and RAF [23]. The RAF is used for large posed and real-world expressions, as the first three do not have large posed expressions. So to check the robustness of our method over large posed expressions, we have used the RAF database.

5.1. Description of Databases. The Extended Cohn-Kanade is the most wide and most popular database used in the facial expression recognition. It contains 593 video sequences, which do vary from 10 to 60 frames with a shift from neutral to other expression. There are a total of 123 subjects who performed different expressions, the ages of the subjects ranging from 18 to 30 years. Out of the 123 subjects, most of them are females. 327 video sequences out of them are categorized into seven expressions. The core reason behind the algorithms not being uniform over CK+ is that it do not provide specific training, validation, and test sets.

The MMI database is laboratory-controlled, and 75 subjects have performed 2900 expressions both video sequences and static images with high resolution, out of which 326 video sequences are obtained from 32 subjects. The MMI database is different from CK+ as it is using both onset, offset, and apex phases. In the sequences, the neutral expression is performed at the start of every sequence and reaches at the peak and, then, return back to the neutral expression. This database is having very challenging conditions, i.e., it is taking care of large interpersonal variations; every subject is performing different nonuniform expression while wearing glasses, moustaches, etc. The Oulu-CASIA database consists of 2880 images from 80 subjects for six expressions; most of them are males between 23 and 58 years. This database is specially designed to tackle the problem of illumination due to the environmental changes. It consists of two different imaging systems; the first one is Near Infrared (NIR), whereas the second one is Visible Light (VIS). There are 3 different variable illumination scenarios: the first one is normal indoor illumination; the second one is used for weak illumination considering the scenario where just the computer display is on, and the third one is having all the lights off, i.e., dark illumination.

The Real-world Affective Faces Database is used, which consists of 29672 great diverse real-world facial images. These images are downloaded from the internet based on the approach of crowdsourcing; 40 annotators are used for independently labelling each image. This database consists of large variability in different subjects' gender, age, ethnicity, varying lighting conditions, head pose, eye gaze, occlusions, and postprocessing operations, which helps us to validate our network over versatile databases.

5.2. Implementation Details. The facial image is first preprocessed using face detection, data augmentation, and illumination normalization (a weighted summation approach has been used in our work for combining histogram equalization and linear mapping) for fine-tuning of the image. The highly uncorrelated data is removed in order to process it further for a high-quality result. Then, the landmark detection is used to detect the key facial points. After that, VGG19 is used as a backbone of the network, where feature maps of $512 \times 7 \times 7$ are obtained after the 1st enhancement module. Then, $256 \times 6 \times 6$ feature maps are obtained from 2×2 convolutional kernels having the stride value of one; those feature maps are further forwarder to primary capsule layers with 8D capsule and 32 convolutional layers. There are 3 routing iterations which are then executed between the primary capsule layers and face capsule layers. Every expression is having a 16D capsule, where all the lower capsules forward information to the above capsule. Then, with 3 fully connected layers,

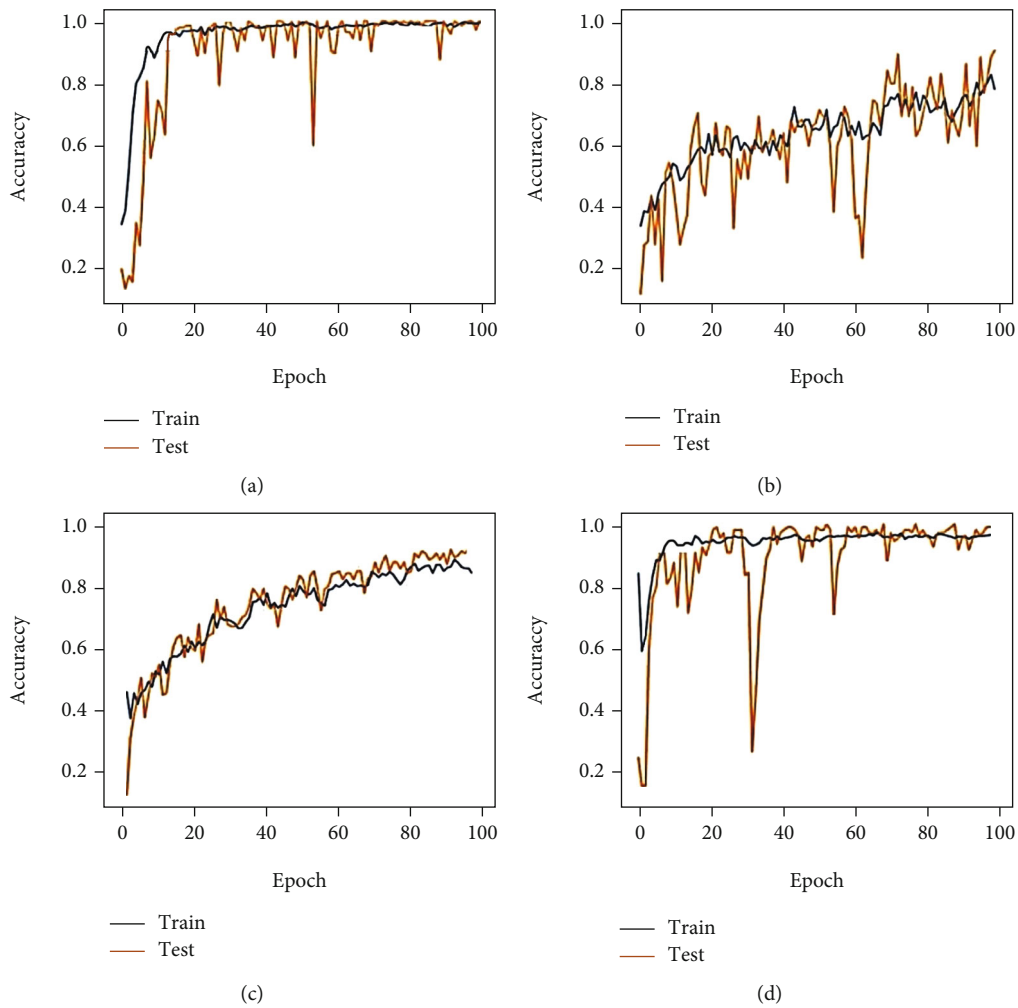


Figure 5: (a) Average accuracy rates on Cohn-Kanade+ Database. (b) Average accuracy rates on MMI Database. (c) Average accuracy rates on Oulu-CASIA Database. (d) Average accuracy rates on CASIA-MFE Database.

we use the squashing function for further classification. Adam optimizer is used for learning with a rate of 0.0001. The value of b^+ is 0.9 (upper boundary) and b^- is 0.1 (lower boundary), respectively. Furthermore, the batch size is set to 16, and the maximum iteration is set to 300. Our whole network training is end-end.

5.3. Discussion. In the Extended Cohn-Kanade database, we take the last frame to three frames and consider the first frame as a neutral expression for data selection. The subjects have been divided in a group of 10, and 10-fold cross-validation is performed. Table 1 shows the average accuracy rates compared with other existing state-of-the-art methods. Our image-based method achieves the highest accuracy of 98.95 percent against sequence-based techniques which extracts the features from a sequence of images or videos. In the MMI database, we take three frames from the middle of each sequence that are associated with peak information and develop a dataset consisting of 624 images. Afterwards, the data augmentation is performed and then distributed among 10 sets. For experimentation, the 10 cross-fold person independent validation is performed using the first frame, i.e., neutral expression, and it takes three peak frames from every frontal sequence. Table 2 shows the dominance in the average accuracy rates compared with other existing methods.

In the Oulu-CASIA database for training and testing, we use the last three frames from every sequence. 10 fold cross-validation is performed just like CK+ in which based on the subject; each fold is completely disjointed with all the remaining folds. Table 3 shows the average accuracy rates, which outperforms all novel methods. It achieves the highest accuracy of 91.2 percent.

Just like other databases in the RAF database, we perform 10-fold cross-validation too. Table 4 shows the average accuracy rates of our method on the RAF database. We first obtained the true positives, false positives, true negatives, and false negatives; then, over 10 folds, we calculated the F1 score and precision per class. Figure 6(a) is showing the per-class precision, and Figure 6(b) is showing the per-class F1 score on the following databases. Whereas, Figure 7 represents the classification results on real-time images.

Table 1: The performance comparison of different approaches on the CK+ database.

Method	Accuracy
LBP TOP [41]	88.99
HOG 3D [16]	91.44
MSR [42]	91.40
STM-Explet [30]	94.19
DTAGN [43]	97.27
3D-CNN-DAP [19]	92.4
NMF-SSCCA [44]	97.3
FER-MPI-SFL (baseline) [45]	98.2
Ours	98.95

Table 2: The performance comparison of different approaches on the MMI database.

Method	Accuracy
LBP TOP [41]	59.51
HOG 3D [16]	60.89
CSPL [46]	73.53
STM-Explet [30]	75.2
DTAGN-Joint [43]	70.3
3D-CNN-DAP [30]	63.4
FER-MPI-SFL (baseline) [45]	83.1
Ours	89.31

Table 3: The performance comparison of different approaches on the Oulu-CASIA database.

Method	Accuracy
LBP TOP [41]	68.1
HOG 3D [16]	70.6
STM-Explet [30]	74.59
Atlases [47]	75.52
DTAGN-Joint [43]	81.46
FN2EN [48]	87.71
PPDN [49]	84.59
FER-MPI-SFL (baseline) [45]	87.39
Ours	91.2

Table 4: The performance comparison of different approaches on the RAF database.

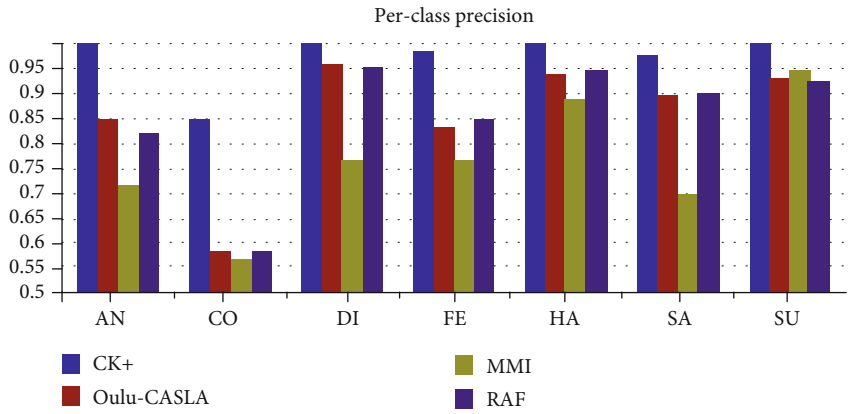
Method	Accuracy
E2E-FC [50]	23.99
AIR [51]	67.37
NAL [52]	84.22
IPA2LT (EM [53] + CNN)	85.30
IPA2(LTNET) (baseline) [54]	86.77
Ours	97.15

5.4. Threats to Validation. There are few factors which can enhance the robustness of facial expression recognition. While validating our approach, there are some limitations of the existing publicly available novel databases. The recognition of the expression with a closed mouth is less accurate as compared with the expression with an open mouth. Considering the agreement of facial expressions by face angles, we noticed that perceived arousal from the frontal face is more than compared with the shift in face angle. Whereas, the happiness disgust with closed mouth and surprise remains unaffected with the face turned away. Furthermore, the effective valence near to the frontal is conveyed more by the full left side profile rather than the full right side profile. It is because of the reason that the left hemiface observes more spontaneous response than the right hemiface. The facial expression analysis can be enhanced by the facial motion information if the image is subtle or degraded. The dynamic neutral expression with blinking of eyes or chewing is also a threat. Moreover, the dwell time is also a key factor; it takes more time over eyes than mouth. However, the dwell time over mouth of happy expression is relatively high. With the increase in the intensity, it can also be noticed that the accuracy is also increased, whereas the dwell time and round trip are decreased. Overall, the response time of females is faster than males even in a low-intensity environment. In the end, it was also concluded that the dwell time of the female eye is more than that of the male.

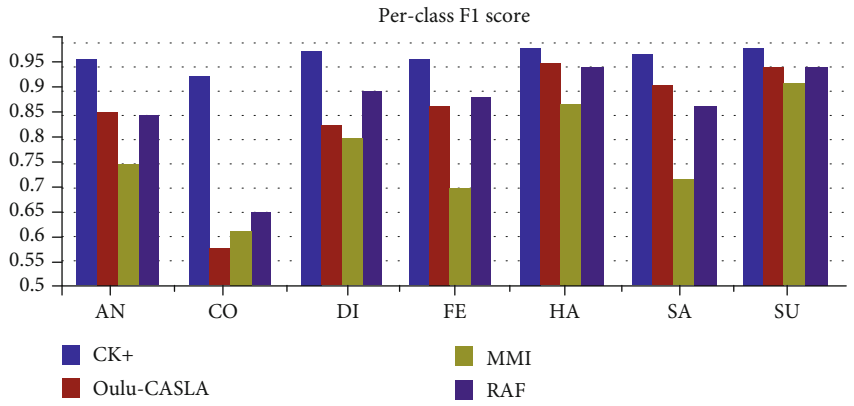
The results and discussion may be presented separately, or in one combined section, and may optionally be divided into headed subsections.

5.5. Potential Applications. The existing architecture has many practical applications in a wide range of areas [55]. We discuss its feasibility in a hospital where patients' health risks are being assessed using their FE's. Such applications are feasible because in a hospital not all patients are attended at all times. For this approach to be effective, hospitals can be installed with sensors in different areas to preprocess patients' data. Fluctuating FE's can be sent to a local sink node, and alarms can be generated based on a specified protocol on a patient-by-patient basis. Another application is the detection of depression among the population in office setup where employees move around frequently. In such wild environments, the head poses of employees can seriously limit the capability of the system to evaluate the facial expressions. Multimodal sensors can play a vital role in such scenarios by extracting the right kind of attributes for better detection. Other possible applications can be in schools and colleges for evaluation of a healthy learning environment for youth, monitoring of babysitters in day-cares across different organizations.

The use of mobile/dynamic multimodal sensing can further enhance the capability of decision-making through enhanced information collection and quality. With dynamic/mobile wireless sensor network (WSN) nodes, it would be possible for nodes to adjust the location to get a better face pose and lower illumination variance. Furthermore, the quality of aggregated data is improved with a fewer number of overall nodes for area coverage. However, mobile multimodal sensing requires complex algorithms for peer communication, area coverage, routing, and security purposes.

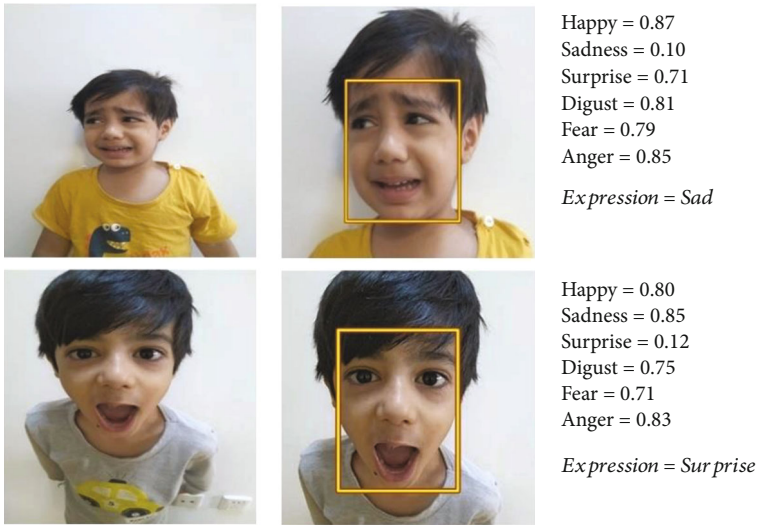


(a) Per-class precision on four databases



(b) Per-class F1 score on four databases

Figure 6: Performance metrics on four databases.



(a) Original image (b) Detected face (c) Average error

Figure 7: Classification results of our proposed method over real-world images.

Multimodal sensors can reduce the overall footprint of the application data by occluding useless background information. Furthermore, redundancy of data can be filtered out using sensor nodes by using local node computation. Apart from that, sensors can be power aided in a local environment thereby creating no power constraints for the resource-constrained WSN nodes.

6. Conclusions

In this paper, we have introduced a state-of-the-art architecture that is robust and effective. A facial image is first preprocessed using different techniques to counter the problems of the excessive background, limitation of data, varying illumination, pose-variation, and occlusion. The facial image is fine-tuned and then forwarded to a dual enhanced capsule network which is capable of handling the spatial transformation. It uses action unit aware mechanism, which helps to locate the active areas which can help in better facial expression recognition. The feature representation ability is enhanced due to multiple convolutional layers, and it helps to capture the key information present in the particular structure of the face.

Different databases have different set of pictures under varying conditions. As a result, the class imbalance is occurred due to the inconsistency in expression annotations. So a cost-sensitive layer can be enhanced for training the deep neural networks. Meanwhile, a powerful deep neural network can be designed having the prior knowledge of change in the local environment, which can be capable of predicting specific parameters and inherently handling and recovering facial occlusions without any intervention. Furthermore, to improve the robustness of the FER, it can be fused with other models. The incorporation with other modalities like depth information from three-dimensional face models, neurosciences, cognitive sciences, infrared images, and physiological data can be a good future research direction.

Data Availability

All the data are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The manuscript is funded with the fund of the Communication University China lab.

References

[1] N. Samadiani, G. Huang, B. Cai et al., "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors*, vol. 19, no. 8, p. 1863, 2019.

[2] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, "Facial expression analysis under partial Occlusion," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–49, 2018.

[3] A. K. Vail, T. Baltrušaitis, L. Pennant, E. Liebson, J. Baker, and L. P. Morency, "Visual attention in schizophrenia: eye contact and gaze aversion during clinical interactions," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 490–497, San Antonio, TX, 2017.

[4] A. Ullah, J. Wang, M. S. Anwar, U. Ahmad, U. Saeed, and Z. Fei, "Facial expression recognition of nonlinear facial variations using deep locality de-expression residue learning in the wild," *Electronics*, vol. 8, no. 12, p. 1487, 2019.

[5] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[6] J. Kumari, R. Rajesh, and K. Pooja, "Facial expression recognition: a survey," *Procedia Computer Science*, vol. 58, pp. 486–491, 2015.

[7] A. Ullah, J. Wang, M. S. Anwar, U. Ahmad, U. Saeed, and J. Wang, "Feature extraction based on canonical correlation analysis using FMEDA and DPA for facial expression recognition with RNN," in *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pp. 418–423, Beijing, China, 2018.

[8] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Identity-free facial expression recognition using conditional generative adversarial network," 2019, https://arxiv.org/abs/1903.08051.

[9] H. Yang, Z. Zhang, and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 294–301, Xi'an, China, 2018.

[10] N. Van Quang, J. Chun, and T. Tokuyama, "Capsulenet for micro-expression recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–7, Lille, France, 2019.

[11] Organization, WHO; others, "Depression: key facts," 2018.

[12] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[13] M. Zhu, D. Shi, and J. Gao, "Branched convolutional neural networks incorporated with jacobian deep regression for facial landmark detection," *Neural Networks*, vol. 118, pp. 127–139, 2019.

[14] Y. Fan, J. C. Lam, and V. O. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *27th International Conference on Artificial Neural Networks*, pp. 84–94, Rhodes, Greece, 2018.

[15] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[16] A. Klaser, M. Marszałek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3d-Gradients," 2008.

[17] W. Cheung and G. Hamarneh, "n-SIFT: n-dimensional scale invariant feature transform," *IEEE Transactions on Image Processing*, vol. 18, no. 9, pp. 2012–2021, 2009.

[18] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 314–321, Santa Barbara, CA, USA, 2011.

[19] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression

analysis," in *12th Asian Conference on Computer Vision*, pp. 143–157, Singapore, Singapore, November 2014, Springer.

[20] B. K. Kim, H. Lee, J. Roh, and S. Y. Lee, "Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 427–434, Seattle, WA, USA, November 2015.

[21] H. W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 443–449, Seattle, WA, USA, November 2015.

[22] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 435–442, Seattle, WA, USA, November 2015.

[23] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2852–2861, Honolulu, HI, USA, July 2017.

[24] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 558–565, Washington, DC, USA, 2017.

[25] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 302–309, Xi'an, China, 2018.

[26] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2168–2177, Salt Lake City, UT, USA, June 2018.

[27] K. Ali, I. Isler, and C. Hughes, "Facial expression recognition using human to animated-character expression translation," 2019, https://arxiv.org/abs/1910.05595.

[28] G. Muhammad, M. Alsulaiman, S. U. Amin, A. Ghoneim, and M. F. Alhamid, "A facial-expression monitoring system for improved healthcare in smart cities," *IEEE Access*, vol. 5, pp. 10871–10881, 2017.

[29] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on multimodal interaction*, pp. 494–501, Istanbul, Turkey, November 2014.

[30] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1749–1756, Columbus, OH, USA, June 2014.

[31] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, 2020.

[32] J. A. Miranda, M. F. Canabal, M. Portela García, and C. Lopez-Ongil, "Embedded emotion recognition: autonomous multimodal affective internet of things," *Proceedings of the cyber-physical systems workshop*, vol. 2208, pp. 22–29, 2018.

[33] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Kauai, HI, USA, 2001.

[34] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Faceness-net: face detection through deep facial part responses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1845–1859, 2018.

[35] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, https://arxiv.org/abs/1511.06434.

[36] M. Valstar, M. Pantic, and I. Patras, "Motion history for facial action detection in video," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, vol. 1, pp. 635–640, The Hague, Netherlands, 2004.

[37] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "EAC-net: deep nets with enhancing and cropping for facial action unit detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 11, pp. 2583–2596, 2018.

[38] P. Lucey, J. F. Cohn, T. Kanade, and J. Saragih, "The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101, San Francisco, CA, USA, 2010.

[39] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*, Amsterdam, Netherlands, 2005.

[40] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[41] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[42] R. Ptucha and A. Savakis, "Manifold based sparse representation for facial understanding in natural images," *Image Vision Computing*, vol. 31, no. 5, pp. 365–378, 2013.

[43] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, Beijing, China, August 2018.

[44] A. Ullah, J. Wang, M. S. Anwar, U. Ahmad, J. Wang, and U. Saeed, "Nonlinear manifold feature extraction based on spectral supervised canonical correlation analysis for facial expression recognition with RRNN," in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–6, Beijing, China, 2018.

[45] W. Wu, Y. Yin, Y. Wang, X. Wang, and D. Xu, "Facial expression recognition for different pose faces based on special landmark detection," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 1524–1529, Beijing, China, 2018.

[46] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2562–2569, Providence, RI, USA, 2012.

[47] Y. Guo, G. Zhao, and M. Pietikäinen, "Dynamic facial expression recognition using longitudinal facial expression atlases," in *Computer Vision – ECCV 2012: 12th European Conference*

on Computer Vision, vol. 7573, pp. 631–644, Florence, Italy, 2012.

[48] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: regularizing a deep face recognition net for expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 118–126, Washington, DC, 2017.

[49] X. Zhao, X. Liang, L. Liu et al., "Peak-piloted deep network for facial expression recognition," in *Computer Vision – ECCV 2016: 14th European Conference*, pp. 425–442, Amsterdam, The Netherlands, 2016.

[50] B. Chen, Y. Wang, G. Wei, J. Li, and B. Ma, "End-to-end trained sparse coding network with spatial pyramid pooling for image classification," *Neural Processing Letters*, vol. 50, no. 3, pp. 2021–2036, 2019.

[51] S. Azadi, J. Feng, S. Jegelka, and T. Darrell, "Auxiliary image regularization for deep cnns with noisy labels," 2015, https://arxiv.org/abs/1511.07069.

[52] J. Goldberger and E. Ben-Reuven, "Training Deep Neural-Networks Using a Noise Adaptation Layer," 2016.

[53] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[54] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 222–237, United Kingdom, 2018.

[55] A. Ullah, J. Wang, M. S. Anwar et al., "Fusion of machine learning and privacy preserving for secure facial expression recognition," *Security and Communication Networks*, vol. 2021, Article ID 6673992, 12 pages, 2021.